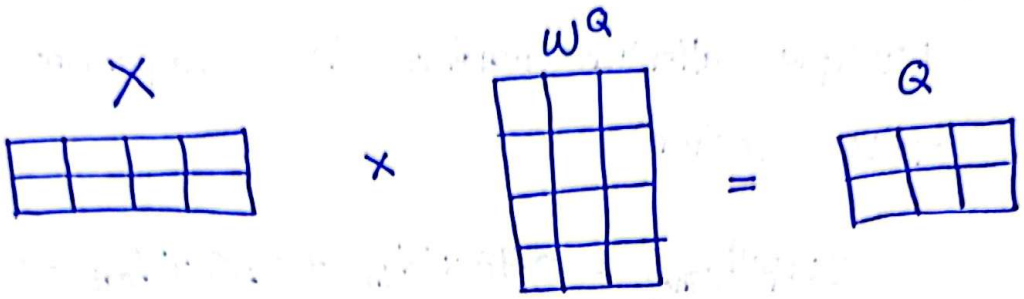
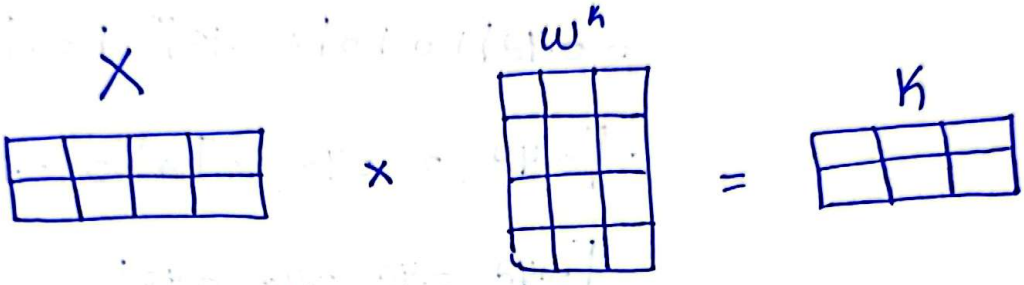
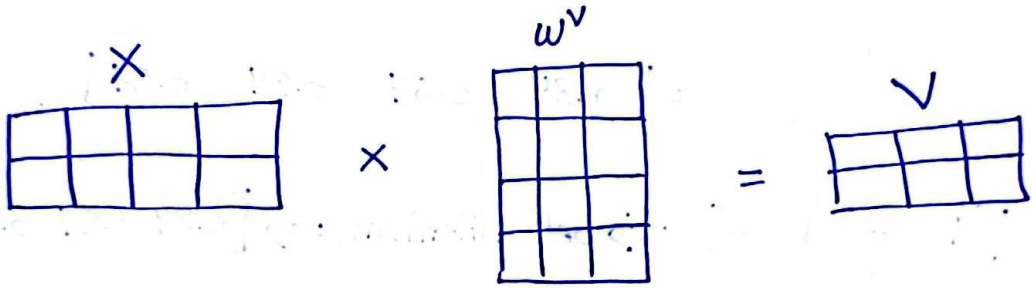
→ Multi-head Attention







$$\text{Softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) V = Z$$

↳ Attention head

Attention head 0	Attention head 1	Attention head N
Q_0 W_0^Q	Q_1 W_1^Q	Q_n W_n^Q
K_0 W_0^K	K_1 W_1^K	K_n W_n^K
V_0 W_0^V	V_1 W_1^V	V_n W_n^V
Z_0	Z_1	Z_n